# Definition

## Project Overview

A key component to any strategic marketing, branding or business growth is market segmentation. The data set analyzed in this Capstone project is a user data collection from a mobile marketplace app for used goods. By segmenting the sellers into multiple groups, the company could provide a better customer support by training support staffs accordingly for each seller group when sellers reach out for assist throughout the process of selling their items.

## Problem Statement

The goal of the project is to segment the sellers into number of groups and investigate the characteristics and uniqueness of each group. The preliminary assessment by the data provider suggests that there are four distinct seller groups: top sellers, business sellers, casual sellers and new sellers.

This capstone project will further investigate the data using various techniques of clustering analysis and will determine the number of unique seller groups based on the given features of the data set.

## Metrics

Silhouette analysis will be used to test the number of clusters after KMeans clustering analysis was conducted on the dataset. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. (scikit learn webpage)

# Analysis

## Data Exploration

The dataset analyzed in this project contains sellers’ activities which are shown as columns (features) in the form of csv file. The descriptions of each column is shown in Table XX.

|  |  |
| --- | --- |
| **Column name** | **Description** |
| id | user ID |
| install\_date | the user install date |
| time\_on\_site | days since the user install date |
| positive\_rating | number of positive ratings the user received as a seller |
| neutral\_rating | number of neutral ratings the user received as a seller |
| negative\_rating | number of negative ratings the user received as a seller |
| listing | number of items the user has listed for sale |
| listing\_gmv | total dollar amount of the listed items from the user |
| sale | number of sales the user made |
| buyers | number of unique buyers of the user’s items |
| gmv | total dollar amount of the user’s sold item |

**Explain about data value of zero**

Though the length of the dataset contains records from over 1.2 million sellers, majority of data contains new users who does not have any selling history using the app (‘sale’ column in the data set is zero). The users that have at least one sell record are only 8.07% of the total number of the sellers. These users are referred as ‘active users’ in the following analysis.

**Distribution of active sellers**

Scatter plots and histograms are shown in Figure XX, which shows very wide distributions of the users across each feature. The plots also represent the skewedness of data toward lower quantities for both listings, sales and total revenues alike. We can assume based on these plots that majority of the sellers collected in the dataset represent new or relatively casual sellers.

Histogram for price or item?

Explain abnormalities in data

Show variance and skewedness of each feature

* If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?
* If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?
* ~~If a dataset is not present for this problem, has discussion been made about the input space or input data for your problem?~~
* Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)

## Exploratory Visualization

* Have you visualized a relevant characteristic or features about the dataset or input data?
* Is the visualization thoroughly analyzed and discussed?
* If a plot is provided, are the axes, title, and datum clearly defined?

## Algorithms and Techniques

PCA, Feature scaling

## Benchmark

# Methodology

## Data Preprocessing

## Implementation

## Refinement

# Results

## Model Evaluation and Validation

Split data 80/20 (at beginning) and train model with 80% and test model with the rest

## Justification

Use F-score?

# Conclusion

## Free-Form Visualization

## Reflection

## Improvement